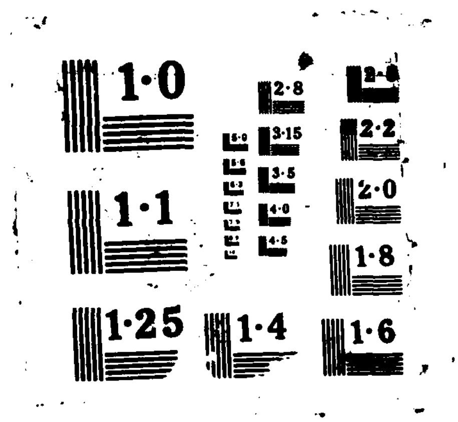
AD-A185 179 ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANALYSIS 1/1
USING BMDP/4M. (U) CORNELL UNIV ITHACA NY MATHEMATICAL
SCIENCES INST W T FEDERER ET AL. 25 MAY 87 TR-87-51
UNCLASSIFIED ARO-23306.86-MA DAAG29-85-C-0018 F/G 12/3 NL


1·0
2·8
3·15
2·2
3·5
2·0
1·1
4·0
4·5
1·8
1·25
1·4
1·6

DTIC FILE COPY MASTER COPY FOR REPRODUCTION PURPOSES (2)

AD-A185 179

REPORT DOCUMENTATION PAGE

1b. RESTRICTIVE MARKINGS	
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARO 23306.86-MA
6a. NAME OF PERFORMING ORGANIZATION Mathematical Sciences Inst. 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION U. S. Army Research Office
6c. ADDRESS (City, State, and ZIP Code) 294 Caldwell Hall; Cornell University Ithaca, New York 14853	7b. ADDRESS (City, State, and ZIP Code) P. O. Box 12211 Research Triangle Park, NC 27709-2211
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U. S. Army Research Office 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DAAL29-85-C-0018
8c. ADDRESS (City, State, and ZIP Code) P. O. Box 12211 Research Triangle Park, NC 27709-2211	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Illustrative Examples of Principal Component Analysis Using BMDP/4M*

12. PERSONAL AUTHOR(S)

W.T. Federer, C.E. McCulloch and N.J. Miles-McDermott

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim Technical	FROM ____ TO ____	May 25, 1987	52

16. SUPPLEMENTARY NOTATION The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

correlations; coefficients; linear combinations; variance-covariance matrix; principal components; annotated computer output

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In order to provide a deeper understanding of the workings of principal components, four data sets were constructed by taking linear combinations of values of two uncorrelated variables to form the X-variates for the principal component analysis. The examples highlight some of the properties and limitations of principal component analysis.

This is part of a continuing project that produces annotated computer output for principal component analysis. The complete project will involve processing four examples on SAS/PRINCOMP, BMDP/4M, SPSS-X/FACTOR, GENSTAT/PCP, and SYSTAT/FACTOR. We show here the results from BMDP/4M, Version 85.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

ARO 23306.86-MA

CORNELL UNIVERSITY

MATHEMATICAL SCIENCES INSTITUTE

TECHNICAL REPORT '87-51

*ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANLAYSIS USING BMDP/4M**

BY

W.T. Federer, C.E. McCulloch and N.J. Miles-McDermott

MAY 1987

Accession For	
NTIS GRA&I	X
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

QUALITY INSPECTED 2

294 Caldwell Hall ■ Ithaca, New York 14853-2602

(607) 255-8005

87 9 9 174

ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANALYSIS USING BMDP/4M*

W. T. Federer, C. E. McCulloch and N. J. Miles-McDermott

BU-929-M February 1987

ABSTRACT

In order to provide a deeper understanding of the workings of principal components, four data sets were constructed by taking linear combinations of values of two uncorrelated variables to form the X-variates for the principal component analysis. The examples highlight some of the properties and limitations of principal component analysis.

This is part of a continuing project that produces annotated computer output for principal component analysis. The complete project will involve processing four examples on SAS/PRINCOMP, BMDP/4M, SPSS-X/FACTOR, GENSTAT / PCP, and SYSTAT / FACTOR. We show here the results from BMDP/4M, Version 85.

* Supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University.

1. INTRODUCTION

Principal components is a form of multivariate statistical analysis and is one method of studying the correlation or covariance structure in a set of measurements on m variables for n observations. For example, a data set may consist of n = 260 samples and m = 15 different fatty acid variables. It may be advantageous to study the structure of the 15 fatty acid variables since some or all of the variables may be measuring the same response. One simple method of studying the correlation structure is to compute the m(m-1)/2 pairwise correlations and note which correlations are close to unity. When a group of variables are all highly inter-correlated, one may be selected for use and the others discarded or the sum of all the variables may be used. When the structure is more complex, the method of principal component analysis (PCA) becomes useful.

In order to use and interpret a principal component analysis, there needs to be some practical meaning associated with the various principal components. In Section 2 we describe the basic features of principal components and in Section 3 we examine some constructed examples using BMDP/4M to illustrate the interpretations that are possible. In Section 4 we summarize our results.

2. BASIC FEATURES OF PRINCIPAL COMPONENT ANALYSIS

PCA can be performed on either the variances and covariances among the m variables or their correlations. One should always check which is being used in a particular computer package program. BMDP/4M, Version 85, can use either the variances and covariances or the correlations but uses the correlations by default. First we will consider analyses using the matrix of variances and covariances. A PCA generates m new variables, the principal components (PCs), by forming linear combinations of the original variables, $\mathbf{X} = (X_1, X_2, \ldots, X_m)$, as follows:

$$PC_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1m}X_m = \mathbf{Xb}_1$$

$$PC_2 = b_{21}X_1 + b_{22}X_2 + \ldots + b_{2m}X_m = \mathbf{Xb}_2$$

$$\vdots$$

$$PC_m = b_{m1}X_1 + b_{m2}X_2 + \ldots + b_{mm}X_m = \mathbf{Xb}_m \quad ,$$

In matrix notation,

$$\mathbf{P} = (PC_1, PC_2, \ldots, PC_m) = \mathbf{X}\ (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m) = \mathbf{XB},$$

and conversely $\mathbf{X} = \mathbf{P}\ \mathbf{B}^{-1}$.

The rationale in the selection of the coefficients, b_{ij}, that define the linear combinations that are the PC_i is to try to capture as much of the variation in the original variables with as few PCs as possible. Since the variance of a linear combination of the Xs can be made arbitrarily large by selecting very large coefficients, the b_{ij} are constrained by convention so that the sum of squares of the coefficients for any PC is unity:

$$\Sigma_{j=1}^{m} b_{ij}^2 = 1 \qquad i = 1,2,\ldots,m \quad .$$

Under this constraint, the b_{1j} in PC_1 are chosen so that PC_1 has maximal variance.

If we denote the variance of X_i by s_i^2 and if we define the total variance as $T = \Sigma_{i=1}^{m} s_i^2$, then the proportion of the variance in the original variables that is captured in PC_1 can be quantified as $var(PC_1)/T$. In selecting the coefficients for PC_2, they are further constrained by the requirement that PC_2 be uncorrelated with PC_1. Subject to this constraint and the constraint that the squared coefficients sum to one, the coefficients b_{2j} are selected so as to maximize $var(PC_2)$. Further coefficients and PCs are selected in a similar manner, by requiring that a PC be uncorrelated with all PCs previously selected and then selecting the coefficients to maximize variance. In this manner, all the PCs are constructed so that they are uncorrelated and so that the first few PCs capture as much variance as possible. The coefficients also have the following interpretation which helps to relate the PCs back to the original variables. The correlation between the i^{th} PC and the j^{th} variable is

$$b_{ij}\sqrt{var(PC_i)}/s_j \quad .$$

After all m PCs have been constructed, the following identity holds:

$$var(PC_1) + var(PC_2) + \ldots + var(PC_m) = T = \Sigma_{i=1}^{m} s_i^2 \quad .$$

This equation has the interpretation that the PCs divide up the total variance of the Xs completely. It may happen that one or more of the last few PCs have variance zero. In such a case, all the variation in the data can be captured by fewer than m variables. Actually, a much stronger result is also true; the PCs can also be used to reproduce the actual values of the Xs, not just their variance. We will demonstrate this more explicitly later.

The above properties of PCA are related to a matrix analysis of the variance-covariance matrix of the Xs, S_x. Let D be a diagonal matrix with entries being the eigenvalues, λ_i , of S_x arranged in order from largest to smallest. Then the following properties hold:

(i) $\lambda_i = \mathrm{var}(PC_i)$

(ii) $\mathrm{trace}(S_x) = \Sigma_{i=1}^{m} s_i^2 = T = \Sigma_{i=1}^{m} \lambda_i = \Sigma_{i=1}^{m} \mathrm{var}(PC_i)$

(iii) $\mathrm{corr}(PC_i, X_j) = \dfrac{b_{ij}\sqrt{\lambda_i}}{s_j}$

(iv) $S_x = B'DB$.

The statements made above are for the case when the analysis is performed on the variance-covariance matrix of the Xs. The correlation matrix could also be used, which is equivalent to performing a PCA on the variance-covariance matrix of the standardized variables,

$$Y_i = \frac{X_i - \bar{X}_i}{s_i}$$

PCA using the correlation martrix is different in these respects:

(i) The total "variance" is m, the number of variables. (It is not truly variance anymore.)

(ii) The correlation between PC_i and X_j is given by

$b_{ij}\sqrt{var(PC_i)} = b_{ij}\sqrt{\lambda_i} = \Lambda_i$. Thus PC_i is most highly correlated with the X_j having the largest coefficient in PC_i in absolute value.

The experimenter must choose whether to use standardized (PCA on a correlation matrix) or unstandardized coefficients (PCA on a variance-covariance matrix). The latter is used when the variables are measured on a comparable basis. This usually means that the variables must be in the same units and have roughly comparable variances. If the variables are measured in different units, then the analysis will usually be performed on the standardized scale, otherwise the analysis may only reflect the different scales of measurement. For example, if a number of fatty acid analyses are made, but the variances, s_i^2, and means, $\bar{X}_i$, are obtained on different bases and by different methods, then standardized variables could be used (PCA on the correlation matrix). To illustrate some of the above ideas, a number of examples have been constructed and these are described in Section 3. In each case, two variables, Z_1 and Z_2, which are uncorrelated, are used to construct X_i. Thus, all the variance can be captured with two variables and hence only two of the PCs will have nonzero variances. In matrix analysis terms, only two eigenvalues will be nonzero. An important thing to note is that in general, **PCA will not recover the original variables** Z_1 and Z_2. Both standardized and nonstandardized computations will be made.

3. EXAMPLES

Throughout the examples we will use the variables Z_1 and Z_2 (with n = 11) from which we will construct $X_1, X_2, \ldots, X_m$. We will perform PCA on the Xs. Thus, in our constructed examples, there will only really be two underlying variables.

Values of Z_1 and Z_2

Z_1	-5	-4	-3	-2	-1	0	1	2	3	4	5
Z_2	15	6	-1	-6	-9	-10	-9	-6	-1	6	15

Notice that Z_1 exhibits a linear trend through the 11 samples and Z_2 exhibits a quadratic trend. They are also chosen to have mean zero and be uncorrelated. Z_1 and Z_2 have the following variance-covariance matrix (a variance-covariance matrix has the variance for the i^{th} variable in the i^{th} row and i^{th} column and the covariance between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column).

Variance-covariance matrix of Z_1 and Z_2

$$\begin{bmatrix} 11 & 0 \\ 0 & 85.8 \end{bmatrix}$$

Thus the variance of Z_1 is 11 and the covariance between Z_1 and Z_2 is zero. Also the total variance is 11 + 85.8 = 96.8. Printed parts of computer output that is repetitive have been omitted in examples 2,3, and 4.

Example 1: In this first example we analyze Z_1 and Z_2 as if they were the data. Thus $X_1 = Z_1$ and $X_2 = Z_2$ and $m = 2$. If PCA is performed on the variance-covariance matrix, then the BMDP output is as follows (BMDP control language for this example and all subsequent examples is in the appendix and the boldface print was typed on computer output to explain the calculation performed):

```
1
BMDP-4M - FACTOR ANALYSIS    -    BMDP Program run

Copyright (C) Regents of University of California.

  BMDP Statistical Software, Inc.
  1964 Westwood Blvd.  Suite 202        Phone (213) 475-5700
  Los Angeles, California  90025          Telex  4992203

Program Version: April 1985
                 (VM/CMS)
Manual  Edition: 1983, 1985 reprint.  State NEWS in the PRINT
                 paragraph for a summary of new features.

  JANUARY 12, 1987  AT 12:14:27

PROGRAM CONTROL INFORMATION

/PROBLEM   TITLE IS 'EXAMPLE 1:PCA ON X1 AND X2'.
/INPUT     VARIABLES ARE 2.
           FORMAT IS FREE.
/VARIABLE  NAMES ARE X1,X2.                        }  Control Language (see appendix for details)
/ROTATE    METHOD=NONE.
/FACTOR    FORM=COVA.
           CONSTANT=0.
/PRINT     COVARIANCE.
           NO CORRELATION.
           NO SHADE.
           CASE=11.
/END

PROBLEM TITLE IS
EXAMPLE 1:PCA ON X1 AND X2

NUMBER OF VARIABLES TO READ IN. . . . . . . . .        2
NUMBER OF VARIABLES ADDED BY TRANSFORMATIONS. .        0
TOTAL NUMBER OF VARIABLES . . . . . . . . . . .        2
NUMBER OF CASES TO READ IN. . . . . . . . . . . TO END
CASE LABELING VARIABLES . . . . . . . . . . . .
```

```
MISSING VALUES CHECKED BEFORE OR AFTER TRANS. . NEITHER
BLANKS ARE. . . . . . . . . . . . . . . . . . . . MISSING
NUMBER OF WORDS OF DYNAMIC STORAGE. . . . . . .   40348
VARIABLES TO BE USED
     1 X1          2 X2

INPUT FORMAT IS
FREE

MAXIMUM LENGTH DATA RECORD IS   80 CHARACTERS.

NUMBER OF VARIABLES TO BE USED. . . . . . . . .    2

WEIGHT VARIABLE . . . . . . . . . . . . . . . .
COVARIANCE MATRIX IS FACTORED.
UNROTATED FACTORS ARE PRINCIPAL COMPONENTS.
NUMBER OF FACTORS IS LIMITED TO THE NUMBER OF EIGENVALUES
     GREATER THAN     0.000
     TIMES THE AVERAGE VARIANCE OF THE VARIABLES
TOLERANCE LIMIT FOR MATRIX INVERSION. . . . . .   0.00010
NO ROTATION IS PERFORMED.

DATA AFTER TRANSFORMATIONS FOR FIRST   11 CASES.
CASES WITH ZERO WEIGHTS AND MISSING DATA NOT INCLUDED.

 C A S E          1          2
 NO. LABEL       X1         X2
----- -------- ---------- ----------
    1               -5         15
    2               -4          6
    3               -3         -1
    4               -2         -6
    5               -1         -9
    6                0        -10
    7                1         -9
    8                2         -6
    9                3         -1
   10                4          6
   11                5         15

NUMBER OF CASES READ. . . . . . . . . . . . . . .    11
```

Variables read in are assigned numbers

Rotations are a technique used in factor analysis, not PCA, and are not considered here.

STATISTICS FOR EACH VARIABLE

VARIABLE	MEAN = $\bar{X}_i$	STANDARD DEVIATION = S_i	COEFFICIENT OF VARIATION = $\frac{S_i}{\bar{X}_i}$	SMALLEST VALUE	SMALLEST STANDARD SCORE $= (X_i - \bar{X}_i)/S_i$	FIRST CASE FOR SMALLEST	LARGEST VALUE	LARGEST STANDARD SCORE	FIRST CASE FOR LARGEST
1 X1	0.00000	3.31662	0.212676E+38	-5.0000	-1.51	1	5.0000	1.51	11
2 X2	0.00000	9.26283	0.417161E+17	-10.0000	-1.08	6	15.0000	1.62	1

CASE NUMBERS ABOVE REFER TO DATA MATRIX BEFORE ANY CASES
HAVE BEEN DELETED DUE TO MISSING DATA.
CASES WITH ZERO WEIGHTS ARE NOT INCLUDED.

COVARIANCE MATRIX = S_{ij}

		X1	X2
		1	2
X1	1	11.000000 = S^2_1	
X2	2	-0.000000 = $S_{12} = S_{21}$	85.800000 = S^2_2

SQUARED MULTIPLE CORRELATIONS (SMC) OF
EACH VARIABLE WITH ALL OTHER VARIABLES

1 X1	0.00000 = r^2_{12}
2 X2	0.00000 = r^2_{21}

CONDITION NUMBER = 7.800 = largest λ_i/smallest λ_i = 85.8/11.0 → Indicates how close the variables are to being perfectly collinear. A very large ratio would indicate a singular or near singular matrix.

EIGENVALUES OF COVARIANCE MATRIX

$85.8000 = \lambda_1$ $11.0000 = \lambda_i$ $\lambda_i = S_i^2$

What BMDP refers to as "factors", is referred to as "principal components" in the text. The reason BMDP uses factors is because the program, although it will do principal component analysis, was written for factor analysis.

COMMUNALITIES OBTAINED FROM 2 FACTORS AFTER 1 ITERATIONS.

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE CORRELATION WITH THE FACTORS. $= r_i^2 . PC_i$

1 X1	1.0000	$= r_1^2;\ PC_1 ,\ PC_2$
2 X2	1.0000	$= r_2^2;\ PC_1 ,\ PC_2$

FACTOR	VARIANCE EXPLAINED	CUMULATIVE PROPORTION OF VARIANCE IN DATA SPACE	IN FACTOR SPACE	CARMINES' THETA
1	$85.8000 = \lambda_1$	0.8864	0.8864	0.8718
2	$11.0000 = \lambda_2$	1.0000	1.0000	

THE VARIANCE EXPLAINED BY EACH FACTOR IS THE EIGENVALUE FOR THAT FACTOR (IF POSITIVE).

TOTAL VARIANCE IS DEFINED AS THE SUM OF THE POSITIVE EIGEN VALUES OF THE COVARIANCE MATRIX.

Total variance = T = 85.8 + 11.0 = 96.8

UNROTATED FACTOR LOADINGS (PATTERN) $= b_i\sqrt{\lambda_i} = A_i$

FOR PRINCIPAL COMPONENTS

		FACTOR $1 = A_1$	FACTOR $2 = A_2$
X1	1	0.000	3.317
X2	2	9.263	0.000
	VP	85.8CO	11.000

THE VP FOR EACH FACTOR IS THE SUM OF THE SQUARES OF THE ELEMENTS OF THE COLUMN OF THE FACTOR LOADING MATRIX CORRESPONDING TO THAT FACTOR. THE VP IS THE VARIANCE EXPLAINED BY THE FACTOR.

1

Note: BMDP does not print out b_i (eigenvectors). To obtain eigenvectors, divide the unrotated factor loadings by $\sqrt{\lambda_i}$

i.e. $b_1 = A_1 / \sqrt{\lambda_1}$

$$= [0 \quad 9.263] / \sqrt{85.8}$$
$$= [0 \quad 1]$$

SORTED FACTOR LOADINGS (PATTERN)

		FACTOR 1	FACTOR 2
X2	2	9.263	0.000
X1	1	0.000	3.317
	VP	85.800	11.000

THE ABOVE FACTOR LOADING MATRIX HAS BEEN REARRANGED SO THAT THE COLUMNS APPEAR IN DECREASING ORDER OF VARIANCE EXPLAINED BY FACTORS. THE ROWS HAVE BEEN REARRANGED SO THAT FOR EACH SUCCESSIVE FACTOR, LOADINGS GREATER THAN 0.5000 APPEAR FIRST. LOADINGS LESS THAN 0.2500 HAVE BEEN REPLACED BY ZERO.

FACTOR SCORE COVARIANCE (COMPUTED FROM FACTOR = $S_{PC_i\ PC_j}$

STRUCTURE AND FACTOR SCORE COEFFICIENTS)

THE DIAGONAL OF THE MATRIX BELOW CONTAINS THE SQUARED

MULTIPLE CORRELATIONS OF EACH FACTOR WITH THE VARIABLES.

		FACTOR 1	FACTOR 2
FACTOR	1	1.000	
FACTOR	2	0.000	1.000

1

ESTIMATED FACTOR SCORES AND MAHALANOBIS DISTANCES (CHI-SQUARE S) FROM EACH CASE TO THE CENTROID OF ALL CASES FOR THE ORIGINAL DATA (2 D.F.) FACTOR SCORES (2 D.F.) AND THEIR DIFFERENCE (0 D.F.). EACH CHI-SQUARE HAS BEEN DIVIDED BY ITS DEGREES OF FREEDOM.

CASE LABEL	NO.	CHISQ/DF 2	CHISQ/DF 2	CHISQ/DF 0	FACTOR 1	FACTOR 2
	1	0.119	2.448	-4.658	1.619	-1.508
	2	0.069	0.937	-1.737	0.648	-1.206
	3	0.037	0.415	-0.755	-0.108	-0.905
	4	0.019	0.392	-0.745	-0.648	-0.603
	5	0.010	0.517	-1.016	-0.972	-0.302
	6	0.007	0.583	-1.152	-1.080	0.000

The first column is computed as $\frac{1}{m}(X - \bar{X})\, D^{-1/2}\, S_x^{-1}\, D^{-1/2}(X - \bar{X})$

which for the first case is

$$0.119 = \frac{1}{2}(-5-0 \quad 15-0) \begin{bmatrix} \frac{1}{\sqrt{11}} & 0 \\ 0 & \frac{1}{\sqrt{85.8}} \end{bmatrix} \begin{bmatrix} \frac{1}{11} & 0 \\ 0 & \frac{1}{85.8} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{11}} & 0 \\ 0 & \frac{1}{\sqrt{85.8}} \end{bmatrix} \begin{bmatrix} -5 - 0 \\ 15 - 0 \end{bmatrix}$$

The second column is computed as $\frac{1}{m}(F - \bar{F})'\,(F - \bar{F})$ where the values in F are from the last two columns, which for the first case is

$$2.448 = \frac{1}{2}(1.619 - 0 \quad -1.508 - 0) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.619 - 0 \\ -1.508 - 0 \end{bmatrix}$$

The third column is computed as m times the difference of column one minus column two which for the first case is

$$-4.658 = 2(0.119 - 2.448).$$

$$\text{Factor } i = (b_{i_1}X_1 + b_{i_2}X_2) / \sqrt{\lambda_i}$$

$$\text{Factor } 1 = (0X_1 + 1X_2) / \sqrt{85.8}$$

Factor 1 = $15 / \sqrt{85.8}$
Ob'n 1
$= 1.619$

7	0.010	0.517	-1.016	-0.972	0.302
8	0.019	0.392	-0.745	-0.648	0.603
9	0.037	0.415	-0.755	-0.108	0.905
10	0.069	0.937	-1.737	0.648	1.206
11	0.119	2.448	-4.658	1.619	1.508

FACTOR SCORE COVARIANCE (COMPUTED FROM FACTOR SCORES)

	FACTOR 1	FACTOR 2
FACTOR 1	1.000	
FACTOR 2	0.000	1.000

This factor score covariance matrix should be identical to the one printed on the previous page. Here it is computed from the PC scores.

We can interpret the results as follows:

1) The first principal component is

$$PC_1 = 0 \cdot X_1 + 1 \cdot X_2 = X_2$$

2) $PC_2 = 1 \cdot X_1 + 0 \cdot X_2 = X_1$

3) $Var(PC_1) = \text{eigenvalue} = 85.8 = Var(X_2)$

4) $Var(PC_2) = \text{eigenvalue} = 11.0 = Var(X_1)$

The PCs may be the same as the Xs whenever the Xs are uncorrelated. Since X_2 has the larger variance, it becomes the first principal component.

If PCA is performed on the correlation matrix, we get slightly different results.

Correlation Matrix of Z_1 and Z_2

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A correlation matrix always has unities along its diagonal and the correlation between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column. PCA in BMDP would yield the following output:

STATISTICS FOR EACH VARIABLE

VARIABLE	MEAN = $\bar{X}_i$	STANDARD DEVIATION = S_i	COEFFICIENT OF VARIATION = $\frac{S_i}{\bar{X}_i}$	SMALLEST VALUE	SMALLEST STANDARD SCORE $= (X_i - \bar{X}_i) / S_i$	FIRST CASE FOR SMALLEST	LARGEST VALUE	LARGEST STANDARD SCORE	FIRST CASE FOR LARGEST
1 X1	0.00000	3.31662	0.212676E+38	-5.0000	-1.51	1	5.0000	1.51	11
2 X2	0.00000	9.26283	0.417161E+17	-10.0000	-1.08	6	15.0000	1.62	1

CASE NUMBERS ABOVE REFER TO DATA MATRIX BEFORE ANY CASES
HAVE BEEN DELETED DUE TO MISSING DATA.
CASES WITH ZERO WEIGHTS ARE NOT INCLUDED.

CORRELATION MATRIX = r_{ij}

r_{ii} is always unity as it is the correlation of X_i with itself.

		X1	X2
		1	2
X1	1	1.000 = r_{11}	
X2	2	-0.000 = r_{12} = r_{21}	1.000 = r_{22}

SQUARED MULTIPLE CORRELATIONS (SMC) OF
EACH VARIABLE WITH ALL OTHER VARIABLES

1 X1	0.00000 = r^2_{12}
2 X2	0.00000 = r^2_{21}

CONDITION NUMBER = 1.000 = largest λ_i / smallest λ_i = 1/1

1

COMMUNALITIES OBTAINED FROM 2 FACTORS AFTER 1 ITERATIONS. = $r^2_{i;\,PC_1,PC_2}$

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE CORRELATION WITH THE FACTORS.

1 X1 1.0000 = $r^2_{1;PC_1,PC_2}$

2 X2 1.0000 = $r^2_{2;PC_1,PC_2}$

FACTOR	VARIANCE EXPLAINED	CUMULATIVE PROPORTION OF VARIANCE IN DATA SPACE	IN FACTOR SPACE	CARMINES' THETA
1	1.0000 = λ_1	0.5000	0.5000	0.0000
2	1.0000 = λ_2	1.0000	1.0000	

UNROTATED FACTOR LOADINGS (PATTERN) = $b_i\sqrt{\lambda_i} = A_i$

FOR PRINCIPAL COMPONENTS

		FACTOR 1 = A_1	FACTOR 2 = A_2
X1	1	0.000	1.000
X2	2	1.000	0.000
	VP	1.000	1.000

$b_i = A_i / \sqrt{\lambda_i}$

$b_1 = A_1 / \sqrt{\lambda_i} = [0 \;\; 1]$

SORTED FACTOR LOADINGS (PATTERN)

		FACTOR 1	FACTOR 2
X2	2	1.000	0.000
X1	1	0.000	1.000
	VP	1.000	1.000

THE DIAGONAL OF THE MATRIX BELOW CONTAINS THE SQUARED MULTIPLE CORRELATIONS OF EACH FACTOR WITH THE VARIABLES.

		FACTOR 1	FACTOR 2
FACTOR	1	1.000	
FACTOR	2	0.000	1.000

CASE LABEL	NO	CHISQ/DF 2	CHISQ/DF 2	CHISQ/DF 0	FACTOR 1	FACTOR 2
	1	2.448	2.448	0.000	1.619	-1.508
	2	0.937	0.937	0.000	0.648	-1.206
	3	0.415	0.415	0.000	-0.108	-0.905
	4	0.392	0.392	0.000	-0.648	-0.603
	5	0.517	0.517	0.000	-0.972	-0.302
	6	0.583	0.583	0.000	-1.080	0.000
	7	0.517	0.517	0.000	-0.972	0.302
	8	0.392	0.392	0.000	-0.648	0.603
	9	0.415	0.415	0.000	-0.108	0.905
	10	0.937	0.937	0.000	0.648	1.206
	11	2.448	2.448	0.000	1.619	1.508

$$Factor_i = (b_{i1}X_1/S_1 + b_{i2}X_2/S_2) / \sqrt{\lambda_i}$$

$$Factor_1 = (0X_1/3.317 + 1X_2/9.26) / 1$$

for case 1,

$$= 15/9.29$$

$$= 1.619$$

FACTOR SCORE COVARIANCE (COMPUTED FROM FACTOR SCORES)

		FACTOR 1	FACTOR 2
FACTOR	1	1.000	

The principal components are again the Xs (standardized Zs) themselves, but the eigenvalues (var(PCs)) are unity since the variables have been standardized first.

Example 2: Let $X_1 = Z_1$, $X_2 = 2Z_1$ and $X_3 = Z_2$. If the analysis is performed on the variance-covariance matrix using BMDP the results are:

STATISTICS FOR EACH VARIABLE

VARIABLE	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	SMALLEST VALUE	SMALLEST STANDARD SCORE	FIRST CASE FOR SMALLEST	LARGEST VALUE	LARGEST STANDARD SCORE	FIRST CASE FOR LARGEST
1 X1	0.00000	3.31662	0.212676E+38	-5.0000	-1.51	1	5.0000	1.51	11
2 X3	0.00000	9.26283	0.417161E+17	-10.0000	-1.08	6	15.0000	1.62	1
3 X2	0.00000	6.63325	0.212676E+38	-10.0000	-1.51	1	10.0000	1.51	11

Note: In this example BMDP prints the results for X_1 through X_3 in the following order: X1 X3 X2. (X_2 is printed last because it is an added variable and computed from X1)

COVARIANCE MATRIX = S_{ij}

		X1	X3	X2
		1	2	3
X1	1	$11.000000 = S_1^2$		
X3	2	$-0.000000 = S_{13} = S_{31}$	$85.800000 = S_3^2$	
X2	3	$22.000000 = S_{12} = S_{21}$	$-.000000 = S_{23} = S_{32}$	$44.000000 = S_2^2$

CORRELATION MATRIX IS SINGULAR. RANK = 2. A GENERALIZED INVERSE IS COMPUTED.

SQUARED MULTIPLE CORRELATIONS (SMC) OF
EACH VARIABLE WITH ALL OTHER VARIABLES

CORRELATION MATRIX IS SINGULAR. IT'S RANK IS 2

1 X1	$1.00000 = r^2_1(2,3)$
2 X3	$0.00000 = r^2_3(1,2)$
3 X2	$0.00000 = r^2_2(1,3)$

SINCE THE CORRELATION MATRIX IS SINGULAR, IT MAY BE DESIRABLE TO REPEAT THE ANALYSIS ELIMINATING THE FOLLOWING VARIABLES.

1 X1

CONDITION NUMBER = 0.2576E+17 = 85.8 / 0 = ∞

1

EIGENVALUES OF COVARIANCE MATRIX

Note: $\Sigma_{i=1}^{m} S_i^2 = \Sigma_{i=1}^{m} \lambda_i$

85.8000 = λ_1 55.0000 = λ_2 0.333067E-14 = λ_3

COMMUNALITIES OBTAINED FROM 2 FACTORS AFTER 1 ITERATIONS.

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE CORRELATION WITH THE FACTORS.

r_i^2; PC_1, PC_2, PC_3

1	X1	1.0000
2	X3	1.0000
3	X2	1.0000

FACTOR	VARIANCE EXPLAINED	CUMULATIVE PROPORTION OF VARIANCE IN DATA SPACE	IN FACTOR SPACE	CARMINES' THETA
1	85.8000 = λ_1	0.6094	0.6094	0.6795
2	55.0000 = λ_2	1.0000	1.0000	

Note: The 3rd factor is dropped because its eigenvalue is 0

UNROTATED FACTOR LOADINGS (PATTERN)

---------------------------------- $= b_i\sqrt{\lambda_i} = A_i$

FOR PRINCIPAL COMPONENTS

		FACTOR 1	FACTOR 2
X1	1	0.000	3.317
X3	2	9.263	0.000
X2	3	0.000	6.633
	VP	85.800	55.000

The 3rd factor loadings would all be zero if they had been printed

$$b_2 = [3.317 \quad 0 \quad 6.633] \;/\; \sqrt{55}$$
$$= [.447 \quad 0 \quad .894]$$

SORTED FACTOR LOADINGS (PATTERN)

		FACTOR 1	FACTOR 2
X3	2	9.263	0.000
X2	3	0.000	6.633
X1	1	0.000	3.317
	VP	85.800	55.000

	FACTOR 1	FACTOR 2
FACTOR 1	1.000	
FACTOR 2	0.000	1.000

CASE LABEL	NO.	CHISQ/DF 2	CHISQ/DF 2	CHISQ/DF 0	FACTOR 1	FACTOR 2
	1	0.041	2.448	-4.813	1.619	-1.508
	2	0.019	0.937	-1.836	0.648	-1.206
	3	0.009	0.415	-0.811	-0.108	-0.905
	4	0.007	0.392	-0.770	-0.648	-0.603
	5	0.007	0.517	-1.022	-0.972	-0.302
	6	0.007	0.583	-1.152	-1.080	0.000
	7	0.007	0.517	-1.022	-0.972	0.302
	8	0.007	0.392	-0.770	-0.648	0.603
	9	0.009	0.415	-0.811	-0.108	0.905
	10	0.019	0.937	-1.836	0.648	1.206
	11	0.041	2.448	-4.813	1.619	1.508

$$\text{Factor}_i = (b_{i1}X_1 + b_{i2}X_2 + b_{i3}X_3) \,/\, \sqrt{\lambda_i}$$

$$\text{Factor}_2 = (.447X_1 + .894X_2 + 0X_3 \,/\, \sqrt{55}$$

for case 1

$$= (.447(-5) + .894(-10)) \,/\, \sqrt{55}$$
$$= -1.508 \text{ within rounding error.}$$

FACTOR SCORE COVARIANCE (COMPUTED FROM FACTOR SCORES)

		FACTOR 1	FACTOR 2
FACTOR	1	1.000	
FACTOR	2	0.000	1.000

Analyzing the Correlation Matrix gives the following results:

STATISTICS FOR EACH VARIABLE

VARIABLE	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	SMALLEST VALUE	SMALLEST STANDARD SCORE	FIRST CASE FOR SMALLEST	LARGEST VALUE	LARGEST STANDARD SCORE	FIRST CASE FOR LARGEST
1 X1	0.00000	3.31662	0.212676E+38	-5.0000	-1.51	1	5.0000	1.51	11
2 X3	0.00000	9.26283	0.417161E+17	-10.0000	-1.08	6	15.0000	1.62	1
3 X2	0.00000	6.63325	0.212676E+38	-10.0000	-1.51	1	10.0000	1.51	11

CORRELATION MATRIX = r_{ij}

		X1 1	X3 2	X2 3
X1	1	1.000 = r_{11}		
X3	2	-0.000 = r_{13} = r_{31}	1.000 = r_{33}	
X2	3	1.000 = r_{21} = r_{12}	-0.000 = r_{23} = r_{32}	1.000 = r_{22}

CORRELATION MATRIX IS SINGULAR. RANK = 2. A GENERALIZED INVERSE IS COMPUTED.

SQUARED MULTIPLE CORRELATIONS (SMC) OF
EACH VARIABLE WITH ALL OTHER VARIABLES

CORRELATION MATRIX IS SINGULAR. IT'S RANK IS 2

1 X1	1.00000
2 X3	0.00000
3 X2	0.00000

SINCE THE CORRELATION MATRIX IS SINGULAR, IT MAY BE DESIRABLE TO REPEAT THE ANALYSIS ELIMINATING THE FOLLOWING VARIABLES.

1	X1
2	X3
3	X2

CONDITION NUMBER = 0.1201E+17 = 2/♦
1

COMMUNALITIES OBTAINED FROM 2 FACTORS AFTER 1 ITERATIONS.

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE CORRELATION WITH THE FACTORS.

1 X1	1.0000
2 X3	1.0000
3 X2	1.0000

FACTOR	VARIANCE EXPLAINED	CUMULATIVE PROPORTION OF VARIANCE IN DATA SPACE	CUMULATIVE PROPORTION OF VARIANCE IN FACTOR SPACE	CARMINES' THETA
1	$2.0000 = \lambda_1$	0.6667	0.6667	0.7500
2	$1.0000 = \lambda_2$	1.0000	1.0000	
3	$0.0000 = \lambda_3$	1.0000		

Note: The 3rd factor is dropped because its eignevalue is 0.

UNROTATED FACTOR LOADINGS (PATTERN)

---------------------------------- $= b_i\sqrt{\lambda_i} = A_i$

FOR PRINCIPAL COMPONENTS

		FACTOR 1	FACTOR 2
X1	1	1.000	0.000
X3	2	0.000	1.000
X2	3	1.000	0.000
	VP	2.000	1.000

The 3rd factor loadings would all be zero if they had been printed

$b_1 = [1 \;\; 0 \;\; 1] / \sqrt{2} = [.707 \;\; 0 \;\; .707]$

SORTED FACTOR LOADINGS (PATTERN)

		FACTOR 1	FACTOR 2
X1	1	1.000	0.000
X2	3	1.000	0.000
X3	2	0.000	1.000
	VP	2.000	1.000

THE DIAGONAL OF THE MATRIX BELOW CONTAINS THE SQUARED
MULTIPLE CORRELATIONS OF EACH FACTOR WITH THE VARIABLES.

	FACTOR 1	FACTOR 2
FACTOR 1	1.000	
FACTOR 2	0.000	1.000

1

CASE LABEL	NO.	CHISQ/DF 2	CHISQ/DF 2	CHISQ/DF 0	FACTOR 1	FACTOR 2
	1	2.448	2.448	0.000	-1.508	1.619
	2	0.937	0.937	0.000	-1.206	0.648
	3	0.415	0.415	0.000	-0.905	-0.108
	4	0.392	0.392	0.000	-0.603	-0.648
	5	0.517	0.517	0.000	-0.302	-0.972
	6	0.583	0.583	0.000	0.000	-1.080
	7	0.517	0.517	0.000	0.302	-0.972
	8	0.392	0.392	0.000	0.603	-0.648
	9	0.415	0.415	0.000	0.905	-0.108
	10	0.937	0.937	0.000	1.206	0.648
	11	2.448	2.448	0.000	1.508	1.619

$$PC_i = (b_{i_1} X_1 / S_1 + b_{i_2} X_2 / S_2 + b_{i_3} X_3 / S_3) / \sqrt{\lambda_i}$$

$$PC_1 = (.707X_1 / 3.317 + .707X_2 / 6.633) / \sqrt{2}$$

for case 1,

$$= (.707(-5) / 3.317 + .707(-10) / 6.633) / \sqrt{2}$$

$= -1.508$ within rounding error

FACTOR SCORE COVARIANCE (COMPUTED FROM FACTOR SCORES)

	FACTOR 1	FACTOR 2
FACTOR 1	1.000	
FACTOR 2	0.000	1.000

There are several items to note in these analyses:

i) There are only two nonzero eigenvalues since given X_1 and X_3, X_2 is computed from X_1.

ii) X_3 is its own principal component since it is uncorrelated with all the other variables.

iii) The sum of the eigenvalues is the sum of the variances, i.e.,

$$11 + 44 + 85.8 = 140.8$$

and

$$1 + 1 + 1 = 3 \ .$$

iv) For the variance-covariance analysis, the ratio of the coefficients of X_1 and X_2 in PC_2 is the same as the ratio of the variables themselves (since $X_2 = 2X_1$).

v) Since there are only two nonzero eigenvalues, only two of the PCs have nonzero variances (are nonconstant).

vi) The coefficients help to relate the variables and the PCs. In the variance-covariance analysis,

$$\mathrm{Corr}(PC_2, X_1) = \frac{(\text{coefficient of } X_1 \text{ in } PC_2)\sqrt{\mathrm{var}(PC_2)}}{\sqrt{\mathrm{var}(X_1)}} = \frac{\Lambda_{12}}{\sqrt{\mathrm{var}(X_1)}}$$

$$= \frac{b_{21}\sqrt{\lambda_2}}{s_1}$$

$$= \frac{.447214\sqrt{55}}{3.16625}$$

$$= 1 \ .$$

In the correlation analysis,

$$\mathrm{Corr}(PC_1, X_1) = b_{11}\sqrt{\lambda_1} = \Lambda_{11} = \text{Component loading for } PC_1,\ X_1$$

$$= .707107\sqrt{2}$$

$$= 1 \ .$$

Thus, in both these cases, the variable is perfectly correlated with the PC.

vii) The Xs can be reconstructed exactly from the PCs with nonzero eigenvalues. For example, in the variance-covariance analysis, X_3 is clearly given by PC_1 . X_1 and X_2 can be recovered via the formulas

$$X_1 = PC_2/\sqrt{5}$$

$$X_2 = 2 \cdot PC_2/\sqrt{5} \quad .$$

As a numerical example,

$$-5 = -11.180/\sqrt{5} \ .$$

Example 3: For Example 3 we use $X_1 = Z_1$, $X_2 = 2(Z_1+5)$, $X_3 = 3(Z_1 +5)$ and $X_4 = Z_2$. Thus X_1, X_2 and X_3 are all created from Z_1. The analyses for the variance-covariance matrix (unstandardized analysis) and correlation matrix (standardized analysis) are given below.

STATISTICS FOR EACH VARIABLE

VARIABLE	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	SMALLEST VALUE	SMALLEST STANDARD SCORE	FIRST CASE FOR SMALLEST	LARGEST VALUE	LARGEST STANDARD SCORE	FIRST CASE FOR LARGEST
1 X1	0.00000	3.31662	0.212676E+38	-5.0000	-1.51	1	5.0000	1.51	11
2 X4	0.00000	9.26283	0.417161E+17	-10.0000	-1.08	6	15.0000	1.62	1
3 X2	10.00000	6.63325	0.663325	0.0000	-1.51	1	20.0000	1.51	11
4 X3	15.00000	9.94987	0.663325	0.0000	-1.51	1	30.0000	1.51	11

COVARIANCE MATRIX = S_{ij}

		X1	X4	X2	X3
		1	2	3	4
X1	1	11.000000			
X4	2	0.000000	85.800000		
X2	3	22.000000	-0.000000	44.000000	
X3	4	33.000000	-0.000000	66.000000	99.000000

CORRELATION MATRIX IS SINGULAR. RANK = 2. A GENERALIZED INVERSE IS COMPUTED.

SQUARED MULTIPLE CORRELATIONS (SMC) OF
EACH VARIABLE WITH ALL OTHER VARIABLES

CORRELATION MATRIX IS SINGULAR. IT'S RANK IS 2

1 X1	1.00000 = $r^2_1(4,2,3)$
2 X4	0.00000
3 X2	1.00000
4 X3	0.00000

SINCE THE CORRELATION MATRIX IS SINGULAR, IT MAY BE DESIRABLE TO REPEAT THE ANALYSIS ELIMINATING THE FOLLOWING VARIABLES.

1 X1
3 X2

1

EIGENVALUES OF COVARIANCE MATRIX = λ_i

154.000   85.8000   0.113243E-13   -0.356271E-14

COMMUNALITIES OBTAINED FROM 2 FACTORS AFTER 1 ITERATIONS.

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE
CORRELATION WITH THE FACTORS. $r^2_{i:\ PC_1, PC_2, PC_3}$

1 X1	1.0000
2 X4	1.0000
3 X2	1.0000
4 X3	1.0000

FACTOR	VARIANCE EXPLAINED	CUMULATIVE PROPORTION OF VARIANCE IN DATA SPACE	IN FACTOR SPACE	CARMINES' THETA
1	154.0000	0.6422	0.6422	0.8143
2	85.8000	1.0000	1.0000	

UNROTATED FACTOR LOADINGS (PATTERN)

FOR PRINCIPAL COMPONENTS

		FACTOR 1	FACTOR 2
X1	1	3.317	0.000
X4	2	0.000	9.263
X2	3	6.633	0.000
X3	4	9.950	0.000
	VP	154.000	85.800

SORTED FACTOR LOADINGS (PATTERN)

		FACTOR 1	FACTOR 2
X3	4	9.950	0.000
X2	3	6.633	0.000
X1	1	3.317	0.000
X4	2	0.000	9.263
	VP	154.000	85.800

		FACTOR 1	FACTOR 2
FACTOR	1	1.000	
FACTOR	2	0.000	1.000

Note: The 3rd and 4th factor loadings are all zero

$$b_1 = [3.317 \quad 0 \quad 6.633 \quad 9.950] \,/\, \sqrt{154}$$
$$= [.267 \quad 0 \quad .535 \quad .802]$$

CASE		CHISQ/DF	CHISQ/DF	CHISQ/DF	FACTOR	FACTOR
LABEL	NO	2	2	0	1	2
	1	0.027	2.448	-4.842	1.508	1.619
	2	0.010	0.937	-1.855	-1.206	0.648
	3	0.004	0.415	-0.821	-0.906	-0.108
	4	0.004	0.392	0.775	0.603	-0.648
	5	0.006	0.517	1.023	-0.302	0.972
	6	0.007	0.583	1.152	0.000	-1.060
	7	0.006	0.517	1.023	0.302	0.972
	8	0.004	0.392	-0.775	0.603	-0.648
	9	0.004	0.415	0.821	0.906	0.108
	10	0.010	0.937	1.855	1.206	0.648
	11	0.027	2.448	4.842	1.508	1.619

FACTOR SCORE COVARIANCE (COMPUTED FROM FACTOR SCORES)

		FACTOR 1	FACTOR 2
FACTOR	1	1.000	
FACTOR	2	0.000	1.000

$$\text{Factor}_i = (b_{i1}(X_1-\bar{X}_1) + b_{i2}(X_2-\bar{X}_2) + b_{i3}(X_3-\bar{X}_3) + b_{i4}(X_4-\bar{X}_4)) \,/\, \sqrt{\lambda_i}$$

$$\text{Factor}_1 = (.267(X_1-0) + 5.35(X_2-10) + .802(X_3-15) + 0(X_4-0)) \,/\, \sqrt{154}$$

for case 1.

$$= (.267(-5) + .535(-10) + .802(-15)) \,/\, \sqrt{154}$$
$$= -1.508$$

STATISTICS FOR EACH VARIABLE

VARIABLE	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	SMALLEST VALUE	SMALLEST STANDARD SCORE	FIRST CASE FOR SMALLEST	LARGEST VALUE	LARGEST STANDARD SCORE	FIRST CASE FOR LARGEST
1 X1	0.00000	3.31662	0.212676E+38	-5.0000	-1.51	1	5.0000	1.51	11
2 X4	0.00000	9.26283	0.417161E+17	-10.0000	-1.08	6	15.0000	1.62	1
3 X2	10.00000	6.63325	0.663325	0.0000	-1.51	1	20.0000	1.51	11
4 X3	15.00000	9.94987	0.663325	0.0000	-1.51	1	30.0000	1.51	11

CORRELATION MATRIX = r_{ij}

		X1	X4	X2	X3
		1	2	3	4
X1	1	1.000			
X4	2	-0.000	1.000		
X2	3	1.000	-0.000	1.000	
X3	4	1.000	-0.000	1.000	1.000

CORRELATION MATRIX IS SINGULAR. RANK = 2. A GENERALIZED INVERSE IS COMPUTED.
SQUARED MULTIPLE CORRELATIONS (SMC) OF
EACH VARIABLE WITH ALL OTHER VARIABLES

CORRELATION MATRIX IS SINGULAR. IT'S RANK IS 2

1 X1	1.00000
2 X4	0.00000
3 X2	1.00000
4 X3	0.00000

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE CORRELATION WITH THE FACTORS.

1	X1	1.0000
2	X4	1.0000
3	X2	1.0000
4	X3	1.0000

FACTOR	VARIANCE EXPLAINED	CUMULATIVE PROPORTION OF VARIANCE IN DATA SPACE	CUMULATIVE PROPORTION OF VARIANCE IN FACTOR SPACE	CARMINES' THETA
1	$3.0000 = \lambda_1$	0.7500	0.7500	0.SSS9
2	$1.0000 = \lambda_2$	1.0000	1.0000	
3	0.0000	1.0000		
4	0.0000			

UNROTATED FACTOR LOADINGS (PATTERN)

FOR PRINCIPAL COMPONENTS

		FACTOR 1	FACTOR 2
X1	1	1.000	0.000
X4	2	0.000	1.000
X2	3	1.000	0.000
X3	4	1.000	0.000
	VP	3.000	1.000

$$b_1 = [1 \quad 0 \quad 1 \quad 1] / \sqrt{3}$$
$$= [.577 \quad 0 \quad .577 \quad .577]$$

SORTED FACTOR LOADINGS (PATTERN)

		FACTOR 1	FACTOR 2
X3	4	1.000	0.000
X2	3	1.000	0.000
X1	1	1.000	0.000
X4	2	0.000	1.000
	VP	3.000	1.000

		FACTOR 1	FACTOR 2
FACTOR	1	1.000	
FACTOR	2	0.000	1.000

CASE LABEL	NO.	CHISQ/DF 2	CHISQ/DF 2	CHISQ/DF 0	FACTOR 1	FACTOR 2
	1	2.448	2.448	0.000	-1.508	1.619
	2	0.937	0.937	0.000	-1.206	0.648
	3	0.415	0.415	0.000	-0.905	-0.108
	4	0.392	0.392	0.000	-0.603	-0.648
	5	0.517	0.517	0.000	-0.302	-0.972
	6	0.583	0.583	0.000	0.000	-1.080
	7	0.517	0.517	0.000	0.302	-0.972
	8	0.392	0.392	0.000	0.603	-0.648
	9	0.415	0.415	0.000	0.905	-0.108
	10	0.937	0.937	0.000	1.206	0.648
	11	2.448	2.448	0.000	1.508	1.619

$$PC_i = (b_{i1}(X_1-\bar{X}_1)/S_1 + b_{i2}(X_2-\bar{X}_2)/S_2 + b_{i3}(X_3-\bar{X}_3)/S_3 + b_{i4}(X_4-\bar{X}_4)/S_4) \,/\, \sqrt{\lambda_i}$$

$$PC_1 = (.577(X_1-0)/3.319 + .577(X_2-10)/6.633 + .577(X_3-15)/9.950 + 0(X_4-0)/9.628$$

for case 1,

$$= (.577(-5)/3.317 + .577(-10)/6.633 + .577(-15)/9.950) \,/\, \sqrt{3}$$
$$= -1.508$$

For the variance-covariance analysis, the coefficients in PC_1 are in the same ratio as their relationship to Z_1. In the correlation analysis X_1, X_2 and X_3 have equal coefficients. In both analyses, as expected, the total variance is equal to the sum of the variances for the PCs. In both cases two PCs, PC_3 and PC_4, have zero variance and are identically zero.

Example 4. In this example we take more complicated combinations of Z_1 and Z_2.

$$X_1 = Z_1$$

$$X_2 = 2Z_1$$

$$X_3 = 3Z_1$$

$$X_4 = Z_1/2 + Z_2$$

$$X_5 = Z_1/4 + Z_2$$

$$X_6 = Z_1/8 + Z_2$$

$$X_7 = Z_2$$

Note that X_1, X_2 and X_3 are colinear (they all have correlation unity) and X_4, X_5, X_6 and X_7 have steadily decreasing correlations with X_1.

The PCAs for the variance-covariance and correlation matrices are given below.

STATISTICS FOR EACH VARIABLE

VARIABLE	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	SMALLEST VALUE	SMALLEST STANDARD SCORE	FIRST CASE FOR SMALLEST	LARGEST VALUE	LARGEST STANDARD SCORE	FIRST CASE FOR LARGEST
1 X1	0.00000	3.31662	0.212676E+38	-5.0000	-1.51	1	5.0000	1.51	11
2 X7	0.00000	9.26283	0.417161E+17	-10.0000	-1.08	6	15.0000	1.62	1
3 X2	0.00000	6.63325	0.212676E+38	-10.0000	-1.51	1	10.0000	1.51	11
4 X3	0.00000	9.94987	0.212676E+38	-15.0000	-1.51	1	15.0000	1.51	11
5 X4	0.00000	9.41010	0.212676E+38	-10.0000	-1.06	6	17.5000	1.86	11
6 X5	0.00000	9.29987	0.418829E+17	-10.0000	-1.08	6	16.2500	1.75	11
7 X6	0.00000	9.27210	0.417578E+17	-10.0000	-1.08	6	15.6250	1.69	11

COVARIANCE MATRIX

		X1	X7	X2	X3	X4	X5	X6
		1	2	3	4	5	6	7
X1	1	11.000000						
X7	2	-0.000000	85.800000					
X2	3	22.000000	-0.000000	44.000000				
X3	4	33.000000	-0.000000	66.000000	99.000000			
X4	5	5.500000	85.800000	11.000000	16.500000	88.550000		
X5	6	2.750000	85.800000	5.500000	8.250000	87.175000	86.487500	
X6	7	1.375000	85.800000	2.750000	4.125000	86.487500	86.143750	85.971875

CORRELATION MATRIX IS SINGULAR. RANK = 2. A GENERALIZED INVERSE IS COMPUTED.

```
SQUARED MULTIPLE CORRELATIONS (SMC) OF
EACH VARIABLE WITH ALL OTHER VARIABLES
--------------------------------------

CORRELATION MATRIX IS SINGULAR. IT'S RANK IS   2
   1 X1          0.00200
   2 X7          1.00000
   3 X2          1.00000
   4 X3          1.00000
   5 X4          1.00000
   6 X5          1.00000
   7 X6          0.00200

SINCE THE CORRELATION MATRIX IS SINGULAR, IT MAY BE DESIRABLE TO REPEAT THE ANALYSIS ELIMINATING THE FOLLOWING VARIABLES.

    2     X7
    3     X2
    4     X3
    5     X4
    6     X5


EIGENVALUES OF COVARIANCE MATRIX
--------------------------------

     347.015          153.794        0.746070E-13     0.113102E-13     0.481525E-14
  -0.203962E-14    -0.104221E-13
```

COMMUNALITIES OBTAINED FROM 2 FACTORS AFTER 1 ITERATIONS.

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE CORRELATION WITH THE FACTORS.

1	X1	1.0000
2	X7	1.0000
3	X2	1.0000
4	X3	1.0000
5	X4	1.0000
6	X5	1.0000
7	X6	1.0000

FACTOR	VARIANCE EXPLAINED	CUMULATIVE PROPORTION OF VARIANCE IN DATA SPACE	CUMULATIVE PROPORTION OF VARIANCE IN FACTOR SPACE	CARMINES' THETA
1	347.0151	0.6929	0.6929	0.9261
2	153.7943	1.0000	1.0000	

UNROTATED FACTOR LOADINGS (PATTERN)

FOR PRINCIPAL COMPONENTS

		FACTOR 1	FACTOR 2
X1	1	0.466	3.284
X7	2	9.171	-1.302
X2	3	0.932	6.567
X3	4	1.398	9.851
X4	5	9.404	0.340
X5	6	9.287	-0.481
X6	7	9.229	-0.891
	VP	347.015	153.794

$$b_1 = [.466\ \ 9.171\ \ .932\ \ 1.398\ \ 9.404\ \ 9.287\ \ 9.229] \,/\, \sqrt{347.015}$$
$$= [.025\ \ .492\ \ .050\ \ .075\ \ .505\ \ .499\ \ .495]$$

SORTED FACTOR LOADINGS (PATTERN)

		FACTOR 1	FACTOR 2
X4	5	9.404	0.340
X5	6	9.287	-0.481
X6	7	9.229	-0.891
X7	2	9.171	-1.302
X3	4	1.398	9.851
X2	3	0.932	6.567
X1	1	0.466	3.284
	VP	347.015	153.794

		FACTOR 1	FACTOR 2
FACTOR	1	1.000	
FACTOR	2	-0.000	1.000

CASE LABEL	NO.	CHISQ/DF 2	CHISQ/DF 2	CHISQ/DF 0	FACTOR 1	FACTOR 2
	1	0.121	2.448	-4.653	1.391	-1.720
	2	0.069	0.937	-1.735	0.472	-1.285
	3	0.037	0.415	-0.755	-0.234	-0.880
	4	0.019	0.392	-0.746	-0.726	-0.506
	5	0.009	0.517	-1.016	-1.004	-0.162
	6	0.007	0.583	-1.152	-1.069	0.152
	7	0.010	0.517	-1.015	-0.920	0.435
	8	0.019	0.392	-0.745	-0.557	0.688
	9	0.037	0.415	-0.755	0.020	0.911
	10	0.068	0.937	-1.738	0.811	1.103
	11	0.116	2.448	-4.662	1.815	1.265

Factor 1, case 1:

$= (.025(-5) + .492(15) + .050(-10) + .075(-15) + .505(12.5) + .499(13.75)$

$+ .495(14.375)) / \sqrt{347.015}$

$= 1.391$

FACTOR SCORE COVARIANCE (COMPUTED FROM FACTOR SCORES)

	FACTOR 1	FACTOR 2
FACTOR 1	1.000	
FACTOR 2	0.000	1.000

STATISTICS FOR EACH VARIABLE

VARIABLE	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	SMALLEST VALUE	SMALLEST STANDARD SCORE	FIRST CASE FOR SMALLEST	LARGEST VALUE	LARGEST STANDARD SCORE	FIRST CASE FOR LARGEST
1 X1	0.00000	3.31662	0.212676E+38	-5.0000	-1.51	1	5.0000	1.51	11
2 X7	0.00000	9.26283	0.417161E+17	-10.0000	-1.08	6	15.0000	1.62	1
3 X2	0.00000	6.63325	0.212676E+38	-10.0000	-1.51	1	10.0000	1.51	11
4 X3	0.00000	9.94987	0.212676E+38	-15.0000	-1.51	1	15.0000	1.51	11
5 X4	0.00000	9.41010	0.212676E+38	-10.0000	-1.06	6	17.5000	1.86	11
6 X5	0.00000	9.29987	0.418829E+17	-10.0000	-1.08	6	16.2500	1.75	11
7 X6	0.00000	9.27210	0.417578E+17	-10.0000	-1.08	6	15.6250	1.69	11

CORRELATION MATRIX

		X1	X7	X2	X3	X4	X5	X6
		1	2	3	4	5	6	7
X1	1	1.000						
X7	2	-0.000	1.000					
X2	3	1.000	-0.000	1.000				
X3	4	1.000	-0.000	1.000	1.000			
X4	5	0.176	0.984	0.176	0.176	1.000		
X5	6	0.089	0.996	0.089	0.089	0.996	1.000	
X6	7	0.045	0.999	0.045	0.045	0.991	0.999	1.000

CORRELATION MATRIX IS SINGULAR. RANK = 2. A GENERALIZED INVERSE IS COMPUTED.

SQUARED MULTIPLE CORRELATIONS (SMC) OF
EACH VARIABLE WITH ALL OTHER VARIABLES

CORRELATION MATRIX IS SINGULAR. IT'S RANK IS 2

1 X1	1.00000
2 X7	1.00000
3 X2	1.00000
4 X3	0.00200
5 X4	1.00000
6 X5	1.00000
7 X6	0.00200

SINCE THE CORRELATION MATRIX IS SINGULAR, IT MAY BE DESIRABLE TO REPEAT THE ANALYSIS ELIMINATING THE FOLLOWING VARIABLES.

1	X1
2	X7
3	X2
5	X4
6	X5

1

COMMUNALITIES OBTAINED FROM 2 FACTORS AFTER 1 ITERATIONS.

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE CORRELATION WITH THE FACTORS.

1	X1	1.0000
2	X7	1.0000
3	X2	1.0000
4	X3	1.0000
5	X4	1.0000
6	X5	1.0000
7	X6	1.0000

FACTOR	VARIANCE EXPLAINED	CUMULATIVE PROPORTION OF VARIANCE IN DATA SPACE	CUMULATIVE PROPORTION OF VARIANCE IN FACTOR SPACE	CARMINES' THETA
1	4.0522	0.5789	0.5789	0.8788
2	2.9478	1.0000	1.0000	
3	0.0000	1.0000		
4	0.0000	1.0000		
5	0.0000	1.0000		
6	0.0000			
7	0.0000			

UNROTATED FACTOR LOADINGS (PATTERN)

FOR PRINCIPAL COMPONENTS

		FACTOR 1	FACTOR 2
X1	1	0.290	0.957
X7	2	0.957	-0.290
X2	3	0.290	0.957
X3	4	0.290	0.957
X4	5	0.993	-0.117
X5	6	0.979	-0.204
X6	7	0.969	-0.247
	VP	4.052	2.948

$$b_1 = [.290 \quad .957 \quad .290 \quad .290 \quad .993 \quad .979 \quad .969] / \sqrt{4.052}$$
$$= [.144 \quad .475 \quad .144 \quad .144 \quad .493 \quad .486 \quad .481]$$

SORTED FACTOR LOADINGS (PATTERN)

		FACTOR 1	FACTOR 2
X4	5	0.993	0.000
X5	6	0.979	0.000
X6	7	0.969	0.000
X7	2	0.957	-0.290
X1	1	0.290	0.957
X3	4	0.290	0.957
X2	3	0.290	0.957
	VP	4.052	2.948

THE DIAGONAL OF THE MATRIX BELOW CONTAINS THE SQUARED
MULTIPLE CORRELATIONS OF EACH FACTOR WITH THE VARIABLES.

	FACTOR 1	FACTOR 2
FACTOR 1	1.000	
FACTOR 2	-0.000	1.000

CASE LABEL	NO.	CHISQ/DF 2	CHISQ/DF 2	CHISQ/DF 0	FACTOR 1	FACTOR 2
	1	2.448	2.448	0.000	1.112	-1.913
	2	0.937	0.937	0.000	0.270	-1.342
	3	0.415	0.415	0.000	-0.366	-0.834
	4	0.392	0.392	0.000	-0.795	-0.389
	5	0.517	0.517	0.000	-1.017	-0.006
	6	0.583	0.583	0.000	-1.033	0.314
	7	0.517	0.517	0.000	-0.842	0.571
	8	0.392	0.392	0.000	-0.445	0.765
	9	0.415	0.415	0.000	0.159	0.897
	10	0.937	0.937	0.000	0.970	0.966
	11	2.448	2.448	0.000	1.987	0.972

Factor 1, case 1:

$$= (.144(-5)/3.317 + .475(15)/9.263 + .144(-10)/6.633 + .144(-15)/9.950$$

$$+ .493(12.5)/9.410 + .486(13.75)/9.30 + .481(14.375)/9.272) / \sqrt{4.052}$$

$$= 1.112$$

FACTOR SCOPE COVARIANCE (COMPUTED FROM FACTOR SCORES)

	FACTOR 1	FACTOR 2
FACTOR 1	1.000	
FACTOR 2	-0.000	1.000

We note several things:

i) In both analyses there are only two eigenvalues that are nonzero indicating that only two variables are needed. This is not readily apparent from the correlation or variance-covariance matrix.

ii) In PC_1, PC_2 and PC_3 where the standardizrd X_1, X_2 and X_3 are the same, they have the same coefficients.

iii) Neither PCA recovers Z_1 and Z_2. The PCAs with nonzero variances have elements of both Z_1 and Z_2 in them, i.e., neither PC_1 or PC_2 is perfectly correlated with one of the Zs.

4. SUMMARY

PCA provides a method of extracting structure from the variance-covariance or correlation matrix. If a multivariate data set is actually constructed in a linear fashion from fewer variables, then PCA will discover that structure. PCA constructs linear combinations of the original data, $\underset{\sim}{X}$, with maximal variance:

$$\underset{\sim}{P} = \underset{\sim}{X}B .$$

This relationship can be inverted to recover the Xs from the PCs (actually only those PCs with nonzero eigenvalues are needed - see example 2). Though PCA will often help discover structure in a data set, it does have limitations. It will not necessarily recover the exact underlying variables, even if they were uncorrelated (Example 4). Also, by its construction, PCA is limited to searching for linear structures in the Xs.

APPENDIX

Control Language

Control Language is typed in upper case and comments are in lower case. Refer to BMDP , Version 1985 for program documentation.

```
Example 1: PCA on Covariance Matrix
/PROBLEM    TITLE IS 'EXAMPLE 1:PCA ON X1 AND X2'.
/INPUT      VARIABLES ARE 2.
            FORMAT IS FREE.
/VARIABLE   NAMES ARE X1,X2.  ⇒  Input variables
/ROTATE     METHOD=NONE.      ⇒  Instructs BMDP not to rotate factors
/FACTOR     FORM=COVA.        ⇒  Specifies PCA on covariance matrix
            CONSTANT=0.       ⇒  Instructs BMDP to restrict factors to those
                                 whose eigenvalues are > 0
/PRINT      COVARIANCE.        }  Instructs BMDP to print the covariance
            NO CORRELATION.    }  matrix and input data
            NO SHADE.
            CASE=11.
/END
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
```

```
Example 1:  PCA on correlation matrix
/PROBLEM    TITLE IS 'EXAMPLE 1:PCA ON X1 AND X2'.
/INPUT      VARIABLES ARE 2.
            FORMAT IS FREE.
/VARIABLE   NAMES ARE X1,X2.
/ROTATE     METHOD=NONE.
/FACTOR     FORM=CORR.           ⇒  Specifies PCA on correlation matrix
            CONSTANT=0.
/PRINT      CASE=11.              }  Instructs BMDP to print the covariance
            NO SHADE.             }  matrix and raw data
/END
```

```
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
```

Example 2: PCA on covariance matrix

```
/PROBLEM    TITLE IS 'EXAMPLE 2:PCA ON X1, X2, AND X3'.
/INPUT      VARIABLES ARE 2.
            FORMAT IS FREE.
/VARIABLE   NAMES ARE X1,X3,X2.
            ADD=1.
/TRANSFORM  X2=2*X1.              ⇒  Computes X2 from X1
/ROTATE     METHOD=NONE.
/FACTOR     FORM=COVA.
            CONSTANT=0.
/PRINT      CASE=11.
            NO SHADE.
            COVARIANCE.
            NO CORRELATION.
/END
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
```

Example 2: PCA on correlation matrix

```
/PROBLEM    TITLE IS 'EXAMPLE 2:PCA ON X1, X2, AND X3'.
/INPUT      VARIABLES ARE 2.
            FORMAT IS FREE.
/VARIABLE   NAMES ARE X1,X3,X2.
            ADD=1.
/TRANSFORM  X2=2*X1.
/ROTATE     METHOD=NONE.
/FACTOR     FORM=CORR.
            CONSTANT=0.
/PRINT      CASE=11.
            NO SHADE.
/END
```

```
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
```

Example 3: PCA on covariance matrix

```
/PROBLEM    TITLE IS 'EXAMPLE 3:PCA ON X1, X2, X3, AND X4'.
/INPUT      VARIABLES ARE 2.
            FORMAT IS FREE.
/VARIABLE   NAMES ARE X1,X4,X2,X3.
            ADD=2.
/TRANSFORM  X2=2*(X1+5).
            X3=3*(X1+5).
/ROTATE     METHOD=NONE.
/FACTOR     FORM=COVA.
            CONSTANT=-1.
/PRINT      COVARIANCE.
            NO CORRELATION.
            NO SHADE.
            CASE=11.
/END
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
```

Example 3: PCA on correlation matrix

```
/PROBLEM    TITLE IS 'EXAMPLE 3:PCA ON X1, X2, X3, AND X4'.
/INPUT      VARIABLES ARE 2.
            FORMAT IS FREE.
/VARIABLE   NAMES ARE X1,X4,X2,X3.
            ADD=2.
/TRANSFORM  X2=2*(X1+5).
            X3=3*(X1+5).
/ROTATE     METHOD=NONE.
/FACTOR     FORM=CORR.
            CONSTANT=-1.
/PRINT      CASE=11.
            NO SHADE.
/END
```

```
-5 15
.
.
.
5 15

Example 4: PCA on covariance matrix
/PROBLEM    TITLE IS 'EXAMPLE 4:PCA ON X1, X2, X3, X4, X5, X6, AND X7'.
/INPUT      VARIABLES ARE 2.
            FORMAT IS FREE.
/VARIABLE   NAMES ARE X1,X7,X2,X3,X4,X5,X6.
            ADD=5.
/TRANSFORM  X2=2*X1.
            X3=3*X1.
            X4=(X1/2)+X7.
            X5=(X1/4)+X7.
            X6=(X1/8)+X7.
/ROTATE     METHOD=NONE.
/FACTOR     FORM=COVA.
            CONSTANT=0.
/PRINT      COVARIANCE.
            NO CORRELATION.
            NO SHADE.
            CASE=11.
/END
-5 15
.
.
.
5 15

Example 4: PCA on correlation matrix
/PROBLEM    TITLE IS 'EXAMPLE 4:PCA ON X1, X2, X3, X4, X5, X6, AND X7'.
/INPUT      VARIABLES ARE 2.
            FORMAT IS FREE.
/VARIABLE   NAMES ARE X1,X7,X2,X3,X4,X5,X6.
            ADD=5.
/TRANSFORM  X2=2*X1.
            X3=3*X1.
            X4=(X1/2)+X7.
            X5=(X1/4)+X7.
            X6=(X1/8)+X7.
/ROTATE     METHOD=NONE.
/FACTOR     FORM=CORR.
            CONSTANT=0.
/PRINT      CASE=11.
            NO SHADE.
/END

-5 15
.
.
.
5 15
```

END
11-87
DTIC